PROBLEM STATEMENTS

Brand Classification

PROBLEM STATEMENT

Indix deals with product data. Most of our data source is the web. We collect information from ecommerce portals, parse them and add it to our index. One of the challenges we face with product data is to identify the brand a particular belongs to.

This episode of hackathon is going to expose you to the challenges in this space. You are given a product dataset which contains just 3 fields, product_title, brand_id and category_id. The problem is to identify the brand_id, using the other features (product_title and category_id). You could treat this as a standard classification problem and arrive at the label (brand_id) for a given input record. The test set would have 2 fields - product_title and category_id.

We would use F-Score to evaluate your classifier's performance.

DATASETS

- classification_train.tsv.gz ORhttp://192.168.0.114/vs_hackathon/classification_train.tsv.gz
- classification_blind_set.tsv.gz ORhttp://192.168.0.114/vs_hackathon/classification_blind_set.tsv.gz

CONTINUOUS EVALUATION

The evaluator will be made available at 3 P.M., Saturday, 2nd April, 2016

BONUS PROBLEM

If you had fun solving the above one and have more time to explore, we would like you take a stab at solving this. Some of the brand_ids are actually duplicates of themselves. This is where your text mining skills would come to the fore. For ex: HP & Hewlett packard are actually the same brand. In our dataset, they have different ids. We would like you to take a stab at identifying aliases.

The blind set for the bonus problem will be published sometime around 7 P.M. , Saturday, 2nd April, 2016